Réseaux de neurones profonds et fusion de données pour la segmentation sémantique d'images aériennes

Nicolas Audebert, Bertrand Le Saux, Sébastien Lefèvre

Réseaux de neurones profonds et fusion de données pour la segmentation sémantique d'images aériennes

Nicolas Audebert[1,2] Bertrand Le Saux[1] Sébastien Lefèvre[2]

[1] ONERA, *The French Aerospace Lab*, F-91761 Palaiseau, France
{nicolas.audebert,bertrand.le_saux}@onera.fr
[2] Univ. Bretagne-Sud, UMR 6074, IRISA, F-56000 Vannes, France
sebastien.lefevre@irisa.fr

Résumé

Ce travail porte sur l'utilisation des réseaux de neurones convolutifs profonds pour la classification dense des images d'observation de la Terre. En particulier, nous entraînons une variante de l'architecture SegNet sur des images aériennes en zone urbaine et étudions différentes stratégies de segmentation sémantique. Nos contributions sont les suivantes : 1) nous étudions la capacité de transfert des caractéristiques apprises sur des images classiques aux images aériennes en utilisant un réseau entièrement convolutif ; 2) nous réalisons la fusion de données hétérogènes (optique et Lidar) en utilisant un nouveau module neuronal dit de correction résiduelle. Nous démontrons la pertinence de ces contributions sur le jeu de données ISPRS Vaihingen 2D Semantic Labeling.

Mots Clé

Apprentissage profond, segmentation sémantique, télédétection, cartographie sémantique, fusion de données.

Abstract

This work investigates the use of deep fully convolutional neural networks (DFCNN) for pixel-wise scene labeling of Earth Observation images. Especially, we train a variant of the SegNet architecture on remote sensing data over an urban area and study different strategies for performing accurate semantic segmentation. Our contributions are the following : 1) we transfer efficiently a DFCNN from generic everyday images to remote sensing images ; 2) we perform data fusion from heterogeneous sensors (optical and laser) using residual correction. Our framework improves state-of-the-art accuracy on the ISPRS Vaihingen 2D Semantic Labeling dataset.

Keywords

Deep learning, semantic segmentation, remote sensing, semantic mapping, data fusion.

1 Introduction

L'introduction des réseaux de neurones entièrement convolutifs par Long et al. [18] a permis d'importants progrès en segmentation sémantique d'images naturelles, sur des jeux de données tels que PASCAL VOC2012 [10] et Microsoft COCO [17]. Si la cartographie automatisée se rapproche de la segmentation sémantique compte-tenu de la nécessité de classifier l'ensemble des pixels d'une image, la nature des images est bien différente de celles traditionnellement manipulées en vision par ordinateur. En effet, les images de télédétection sont acquises par avion ou par satellite et ne se limitent que rarement aux trois canaux optiques Rouge-Vert-Bleu (RVB). Les objets considérés présentent alors tous une même perspective dans un plan 2D, ce qui réduit l'information de profondeur. Par ailleurs, contrairement aux images de type PASCAL VOC, chaque pixel de l'image possède une sémantique, car il n'existe pas de distinction claire entre un premier plan et un arrière-plan.

Les premières tentatives de cartographie automatisée utilisant l'apprentissage profonds s'intéressaient aux approches par patch [15]. Les images étaient pré-segmentées et chaque région était classifiée par un réseau de neurones. L'introduction de réseaux entièrement convolutifs a permis de s'affranchir de cette pré-segmentation en l'intégrant dans la phase d'apprentissage, le réseau prédisant alors directement une carte de même résolution que l'image d'entrée [20].

Nous présentons ici une méthode de segmentation sémantique d'images aériennes pour la cartographie en zone urbaine en utilisant le jeu de données ISPRS [28] et un réseaux de neurones entièrement convolutifs (SegNet). Puis, à partir de cette approche, nous élaborons une stratégie de fusion de données hétérogènes par correction résiduelle.

2 Contexte

2.1 Segmentation sémantique

En vision par ordinateur, la segmentation sémantique est la tâche définie comme l'assignation d'une classe à chaque région cohérente d'une image. Celle-ci peut être réalisée notamment en classifiant chaque pixel de l'image. De nombreux travaux récents ont montré l'efficacité des réseaux de neurones entièrement convolutifs dans ce cadre, en suivant les principes introduits par Long et al. [18]. En densi-

fiant les dernière couches d'un réseau classifieur classique, il est possible d'obtenir non plus un vecteur de probabilités mais des cartes de chaleur indiquant la probabilité des différentes classes en tout point. Plusieurs adaptations ultérieures ont été proposées, comme le retrait des couches de sous-échantillonnage [7] et la dilatation des couches convolutives [35] pour conserver la résolution. Plusieurs modèles s'inspirent en outre des auto-encodeurs convolutifs [36] et présentent une architecture symétrique encodeur/décodeur, tels que DeconvNet [24] et SegNet [3]. L'intégration de modèles structurés tels que les champs de Markov aléatoires ont également été étudiés [38, 1]. Enfin, l'introduction de nouveaux principes, comme l'apprentissage par résidu [13, 32] et les réseaux de neurones récurrents [33], a permis également d'améliorer l'état-de-l'art sur plusieurs jeux de données de référence.

2.2 Cartographie automatisée à partir d'images d'observation de la Terre

L'apprentissage profond pour l'observation de la Terre est un domaine particulièrement actif depuis les premiers réseaux de neurones convolutifs pour l'extraction des routes [21]. L'utilisation de tels réseaux, pré-entraînés sur des images naturelles, comme générateurs de caractéristiques a été démontrée comme plus efficace que l'utilisation de caractéristiques expertes pour la classification par machine à vecteur de support (SVM) [26]. De cette façon, Lagrange et al. [15] ont obtenu d'excellents résultats en cartographie en combinant superpixels et caractéristiques profondes dans le cadre du *Data Fusion Contest* 2015 [5]. L'utilisation de réseaux de neurones convolutifs multi-échelles [37] puis entièrement convolutifs [20] ont permis d'améliorer cette première approche. En effet, ces réseaux entièrement convolutifs (*Fully-Convolutional Networks* - FCN) apprennent non seulement la sémantique des pixels individuels, mais également les structures spatiales qui les relient, les rendant donc particulièrement adaptés à la cartographie à partir d'images aériennes.

3 Méthode proposée

3.1 Pré-traitement des données

Les images aériennes à haute résolution sont généralement trop grandes pour être traitées en une seule passe par un réseau de neurones convolutifs (*Convolutional Neural Network* - CNN). Par exemple, la taille moyenne d'une tuile du jeu de données ISPRS Vaihingen est 2493×2063, tandis que la majorité des CNN se limitent à une entrée de 256×256 pixels. Compte-tenu des limitations de la quantité de mémoire des GPU, nous divisons les images haute résolution en les découpant par fenêtre glissante. Dans le cas où le pas de la fenêtre glissante est inférieure à sa taille, les prédictions sont moyennées sur les pixels sujets au recouvrement. Ceci permet d'affiner les prédictions le long des bords de la fenêtre et de lisser les discontinuités qui pourraient apparaître.

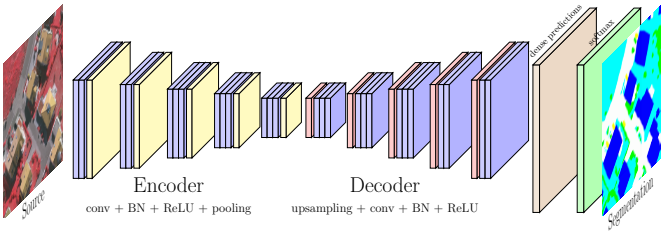


FIGURE 1 – Réseau de neurones entièrement convolutif type SegNet

Notre objectif étant d'adapter des réseaux validés par la communauté vision pour des données de télédétection, les modèles à notre disposition sont originellement conçus pour des images RVB à 3 canaux. Or le jeu de données ISPRS comprend des acquisitions optiques infrarouge-rouge-vert (IRRV) de la ville de Vaihingen. Ces canaux présentent une dynamique similaire et ont été acquises par le même capteur, donc les données IRRV seront traitées comme des images RVB. Le jeu de données inclut également des données acquises par laser (Lidar), desquelles a été dérivé un modèle de surface (DSM). Nous utilisons également le modèle de surface normalisé (NDSM) de [11]. Ces données indiquent la hauteur mesurée par laser en tout point des images optiques. Enfin, nous calculons l'indice *Normalized Divergence Vegetation Index* (NDVI) à partir des canaux infrarouge et rouge, réputé pour être un bon indicateur de végétation :

$$NDVI = \frac{IR - R}{IR + R} \,. \tag{1}$$

Pour chaque image IRRV, nous construisons donc une image composite agrégeant le DSM, le NDSM et le NDVI. Bien qu'hétérogènes, ces données sont complémentaires et combinent des informations pertinentes pour toutes les classes. En effet, la hauteur dérivée du DSM permet de distinguer aisément une route d'un toit ou un arbuste d'un arbre. Par ailleurs, l'information de hauteur est entièrement absente de l'image optique IRRV. Par conséquent, nous souhaitons étudier comment fusionner ces deux sources d'information pour tirer parti au maximum de leur complémentarité.

3.2 Architecture du réseau de neurones

SegNet. De nombreuses architectures de FCN sont disponibles pour la segmentation sémantique. Nous retenons le modèle SegNet [3] (cf. Figure 1) qui présente un équilibre satisfaisant entre précision de la classification et temps de calcul. L'architecture de SegNet est symétrique et permet de replacer précisément les caractéristiques abstraites aux bonnes localisations spatiales. En outre, les résultats préliminaires avec les modèles FCN [18] et DeepLab [7] n'ont pas permis de constater d'améliorations significatives. Toutefois, nous soulignons que nos contributions ne sont pas spécifiques au modèle SegNet et peuvent être adaptées à n'importe quelle autre architecture.

SegNet présente une architecture encodeur-décodeur conçue sur la base des couches de convolution du modèle VGG-16 [6, 30]. L'encodeur est une succession de couches convolutives suivies par une normalisation par batch [14] et des fonctions de transfert non linéaires. Chaque bloc de 2 ou 3 convolutions est suivi par une couche de sous-échantillonnage de pas égal à 2.
Le décodeur est une symétrie de l'encodeur et possède le même nombre de convolutions et le même nombre de blocs. Les réductions de dimensions sont remplacées par des sur-échantillonnages. Ceux-ci replacent les valeurs des activations intermédiaires aux indices ("$argmax$") calculés lors du sous-échantillonnage. Par exemple, la première couche de sous-échantillonnage calcule le masque des activations maximales et le transfère directement à la dernière couche de sur-échantillonnage. Les avant-dernières activations sont alors replacées aux positions ainsi transférées et le reste des cartes d'activation sont remplies par des zéros. Ces cartes d'activations éparses sont ensuite densifiées grâce aux convolutions successives.
L'encodeur étant calqué sur VGG-16, ses poids sont initialisés à partir de ce même CNN pré-entraîné sur le jeu de données ImageNet. Les poids du décodeur sont eux initialisés aléatoirement en utilisant la stratégie décrite dans [12].
Si N désigne le nombre de pixels dans une image d'entrée, k le nombre de classes, et pour un pixel donné i, si y^i représente son étiquette et $(z_1^i, \ldots, z_k^i)$ son vecteur de probabilité issu de SegNet, nous cherchons alors à minimiser la fonction de coût logistique multinomiale moyennée sur toute l'image :

$$loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k} y_j^i \log\left(\frac{\exp(z_j^i)}{\sum_{l=1}^{k}\exp(z_l^i)}\right) . \quad (2)$$

3.3 Fusion de données hétérogènes par correction résiduelle

Les images optiques à 3 canaux ne sont qu'un type possible de données de télédétection. L'imagerie multispectrale contient généralement de 4 à 12 bandes, tandis que l'imagerie hyperspectrale peut contenir plus de 200 canaux différents. En outre, des images Lidar ou Radar peuvent également être disponibles. Comme détaillé dans la Section 3.1, les images IRRV du jeu de données ISPRS sont complétées par un DSM, un NDSM et l'indice NDVI. Nous nous intéressons donc à évaluer s'il est posible : 1) de construire un deuxième SegNet capable de travailler sur ce second ensemble de données, 2) de combiner ces deux réseaux pour améliorer les cartes obtenues.
Une approche naïve à la fusion de données serait de concaténer les 6 canaux (IR/R/V et DSM/NDSM/NDVI) et d'apprendre une architecture SegNet sur cette entrée. Toutefois, l'expérience montre que cette méthode n'améliore pas la précision de la segmentation par rapport à une architecture simplement basée sur IRRV. Inspirés par la fusion multimodale audio-vidéo de [22] et la fusion RVB-profondeur de [9], nous proposons une approche de fusion tardive, travaillant sur les prédictions plutôt que sur les données brutes. Ainsi, nous explorons deux stratégies : 1) le calcul de la prédiction moyenne à partir des cartes de probabilités (Figure 2a), 2) l'apprentissage d'une fusion par réseau de neurones (Figure 2b). Cette dernière utilise un réseau correcteur capable d'apprendre à partir des activations des deux réseaux parallèles comment corriger la prédiction moyenne pour raffiner les cartes.
Dans la continuation de l'apprentissage par résidu [13], nous proposons un module neuronal de correction résiduelle. Prenant en entrée les activations intermédiaires des deux SegNet parallèles, ce réseau apprend la correction à appliquer sur la prédiction moyenne, comme illustré par Figure 3.
Si P_0 représente le tenseur de la vérité terrain et P_i les cartes de probabilité du $i^{\text{ème}}$ SegNet, nous avons :

$$P_i = P_0 + \epsilon_i \text{ avec } |\epsilon_i| \ll |P_i| \quad (3)$$

ϵ_i est un terme d'erreur faible tant que la prédiction P_i est suffisamment précise. Le réseau doit idéalement apprendre à estimer l'erreur et la confiance à accorder à chaque prédiction afin de combiner de manière pertinente les deux flux.
Soit R le nombre de sorties sur lesquelles s'applique la correction résiduelle. Nous prédisons P', la somme de la prédiction moyenne et du terme correctif c :

$$P' = P_{avg} + c = \frac{1}{R}\sum_{i=1}^{R} P_i + c = P_0 + \frac{1}{R}\sum_{i=1}^{R}\epsilon_i + c \quad (4)$$

Le module de correction résiduelle étant optimisé pour minimiser la fonction de coût, nous avons :

$$\|P' - P_0\| \to 0 \quad (5)$$

cette contrainte se traduit sur c et ϵ_i par :

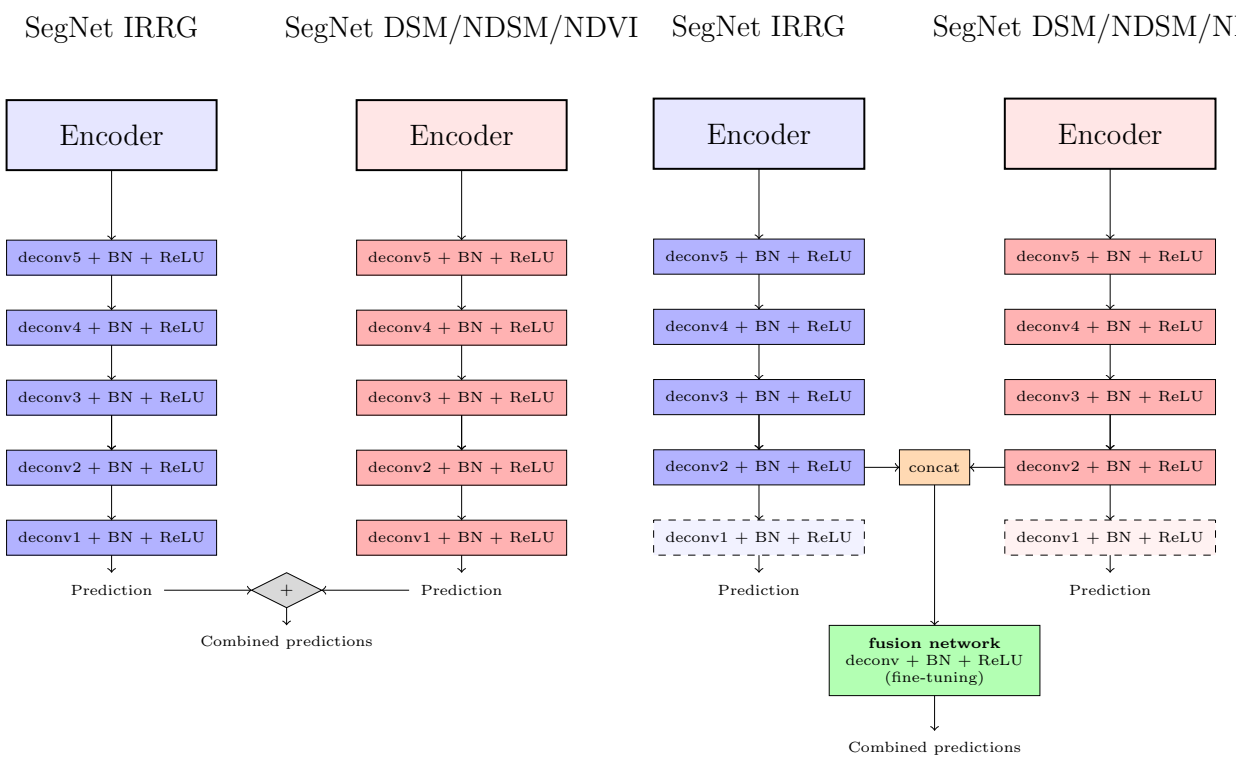
$$\|\frac{1}{R}\sum_{i=1}^{R}\epsilon_i - c\| \to 0 \quad (6)$$

Ceci peut s'interpréter comme la modélisation de l'erreur moyenne à partir des cartes d'activation. En effet, lors de l'apprentissage, la vérité terrain P_0 est connue et la correction résiduelle apprend à inférer $\sum_{i=1}^{R}\epsilon_i$. Le concept d'apprentissage par résidu est bien adapté à cette idée de correction d'erreur, car le résidu est supposé avoir une faible amplitude par rapport au signal principal de contournement (qui est une simple fonction identité).

4 Expériences

4.1 Conditions expérimentales

Nous traitons chaque tuile du jeu de données ISPRS Vaihingen par une fenêtre glissante de dimensions 128×128 et



(a) Fusion par moyenne.

(b) Fusion par apprentissage.

FIGURE 2 – Stratégies de fusion pour notre modèle de SegNet à deux flux.

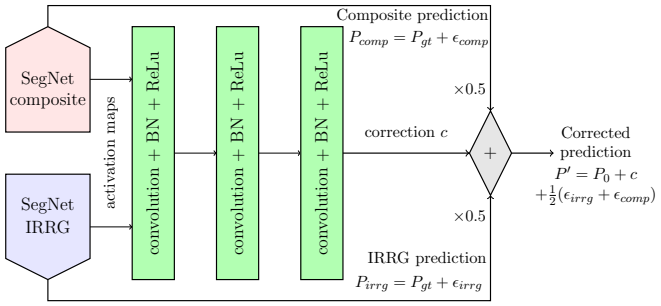


FIGURE 3 – Réseau de fusion apprenant la correction résiduelle à appliquer à la moyenne des prédictions hétérogènes.

un pas de 32 pixels. Nous entraînons séparément les deux SegNet pendant 10 passes sur les données d'apprentissage avec un taux d'apprentissage de 0,1, divisé par 10 après 5 passes. Le SegNet utilisé est la variante multi-contexte incluant plusieurs tailles de noyaux de convolutions dans la dernière couche utilisée dans [2], notée SegNet++. Le modèle de correction résiduelle est appris dans un second temps pendant une passe complète, les poids des SegNet étant alors fixés à leurs valeurs entraînées.

Pour les expériences préliminaires, notre méthode est évaluée uniquement sur les données pour lesquelles une vérité terrain est disponible, que nous divisons en deux sous-ensembles : apprentissage et validation. Pour comparer notre méthode à l'état-de-l'art, nous entraînons ensuite notre modèle sur l'ensemble du jeu de données (apprentissage + validation) avec les mêmes hyperparamètres. Nous soumettons enfin nos résultats sur le jeu de données de test au serveur d'évaluation de l'ISPRS, dont la vérité terrain nous est inconnue.

TABLE 1 – Résultats des différentes stratégies d'initialisation sur le jeu de validation.

Init.	Aléatoire	VGG-16			
Ratio $\frac{lr_e}{lr_d}$	1	1	0.5	0.1	0
Qualité	87.0%	87.2%	**87.8%**	86.9%	86.5%

TABLE 2 – Résultats sur le jeu de validation.

Modèle/Pas (px)	128	64	32
SegNet IRRG	87.8%	88.3%	88.8%
Fusion (moyenne)	88.2%	88.7%	89.1%
Fusion (correction)	88.6%	89.0%	89.5%

4.2 Résultats

Comme démontré dans [26], les filtres convolutifs appris sur des images naturelles peuvent être efficacement transférées pour travailler sur des images aériennes. Toutefois,

TABLE 3 – Résultats du ISPRS 2D Semantic Labeling Challenge Vaihingen.

Méthode	imp surf	building	low veg	tree	car	Qualité
Stair Vision Library ("SVL_3")[11]	86,6%	91,0%	77,0%	85,0%	55,6%	84,8%
RF + CRF ("HUST")[27]	86,9%	92,0%	78,3%	86,9%	29,0%	85,9%
CNN ensemble ("ONE_5")[4]	87,8%	92,0%	77,8%	86,2%	50,7%	85,9%
FCN ("UZ_1")[31]	89,2%	92,5%	81,6%	86,9%	57,3%	87,3%
FCN ("UOA")[16]	89,8%	92,1%	80,4%	88,2%	82,0%	87,6%
CNN + RF + CRF ("ADL_3")[25]	89,5%	93,2%	82,3%	88,2%	63,3%	88,0%
FCN ("DLR_2")[20]	90,3%	92,3%	82,5%	89,5%	76,3%	88,5%
FCN + RF + CRF ("DST_2")[29]	90,5%	93,7%	83,4%	89,2%	72,6%	89,1%
FCN + CRF + frontières ("DLR_10")[19]	92.3%	95.2%	84.1%	90.0%	79.3%	90.3%
SegNet++ (multi-kernel)[2]	**91,5**%	94,3%	82,7%	89,3%	**85,7**%	89,4%
SegNet++ (multi-kernel + fusion)	91,0%	**94,5**%	**84,4**%	**89,9**%	77,8%	**89,8**%

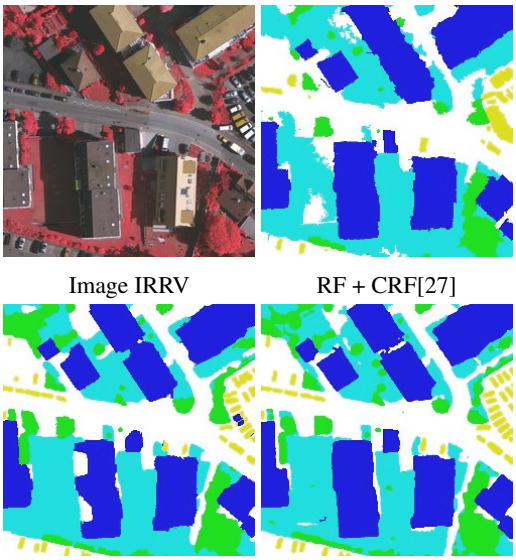


Image IRRV | RF + CRF[27]

"DLR" (FCN)[20] | **SegNet++**

FIGURE 4 – Comparaison des segmentations obtenues sur un extrait du jeu de test ISPRS Vaihingen.
Légende = blanc : routes, bleu : bâtiments, cyan : végétation basse, vert : arbres, Jaune : véhicules.

nous suggérons que de telles images présentent une régularité et une structure spatiale particulière liée au point de vue vertical utilisé. Ainsi, il peut exister un intérêt à laisser ces filtres pré-calculés être modifiés librement lors de l'optimisation du réseau. Pour évaluer cette hypothèse, nous comparons différents taux d'apprentissage pour l'encodeur (lr_e) et le décodeur (lr_d). Nous testons quatre stratégies :

- même variabilité : $lr_d = lr_e$, $lr_e/lr_d = 1$,
- faible variabilité de l'encodeur : $lr_d = 2 \times lr_e$, $lr_e/lr_d = 0,5$,
- très faible variabilité de l'encodeur : $lr_d = 10 \times lr_e$, $lr_e/lr_d = 0,1$,
- absence de rétropropagation de l'erreur sur l'encodeur (aucune variabilité) : $lr_e = 0$, $lr_e/lr_d = 0$.

Ces valeurs seront comparées avec l'initialisation aléatoire de tous les paramètres (encodeur et décodeur), correspondant à l'apprentissage d'un SegNet sans aucun pré-entraînement (et donc aucun transfert de connaissances).
Notre meilleur modèle améliore l'état-de-l'art sur le jeu de données ISPRS Vaihingen (cf. Tableau 3) [1]. La Figure 4 illustre une comparaison qualitative entre différentes méthodes. Les métriques utilisées sont la qualité, définie comme le taux de bonne classification, et les scores F1 par classe :

$$qualité = \frac{tp + tn}{tp + tn + fp + fn} \tag{7}$$

$$F1_i = 2\,\frac{précision_i \times rappel_i}{précision_i + rappel_i} \tag{8}$$

$$rappel_i = \frac{tp_i}{C_i},\ précision_i = \frac{tp_i}{P_i}\ , \tag{9}$$

avec tp, tn, fp, fn respectivement le nombre de vrais positifs, vrais négatifs, faux positifs et faux négatifs, tp_i le nombre de vrais positifs de la classe i, C_i le nombre de pixels appartenant à la classe i et P_i le nombre de pixels attribués à la classe i par le modèle. Ces métriques sont calculées en ignorant un rayon de 3 pixels autour des bordures afin de tenir compte d'éventuelles imprécisions dans la vérité terrain.
La meilleure méthode précédant notre soumission utilisait une combinaison de FCN et de caractéristiques expertes, tandis que la nôtre n'utilise que l'apprentissage statistique. La meilleure méthode précédente utilisant uniquement un FCN ("DLR_1") atteint 88,4%, ce que nous améliorons de 1,4%. Les précédentes méthodes utilisant les CNN atteignent 85,9% ("ONE_5"[4]) et 86,1% ("ADL_1"[25]). Notre méthode obtient des résultats supérieurs, sans recourir à des caractéristiques expertes ou à des post-traitement structurés comme les Champs Aléatoires Conditionnels (*Conditional Random Fields* - CRF). [19] obtient des résultats supérieurs en utilisant un ensemble de trois réseaux de

1. Résultats détaillés : `http://www2.isprs.org/vaihingen-2d-semantic-labeling-contest.html`

neurones parallèles pour la fusion. Toutefois, cette méthode utilise un apprentissage multi-tâche incluant la prédiction de frontières comme régularisation et utilise un NDSM corrigé manuellement comme donnée additionnelle. Sans ce NDSM corrigé, le taux de bonne classification tombe à 89,4%, un résultat inférieur à notre méthode.

4.3 Discussion

Recouvrement de la fenêtre glissante. Autoriser un recouvrement dans la fenêtre glissante augmente le temps d'inférence mais accroît également la précision du modèle, comme détaillé dans le Tableau 2. En effet, en divisant le pas par 2, le nombre d'imagettes à traiter est multiplié par 4. Cependant, moyenner plusieurs prédictions sur une même région permet de corriger des artefacts de classification, notamment le long des bords où le contexte spatial est manquant. L'expérience semble indiquer qu'un pas de 32px (75% de recouvrement) est suffisamment rapide pour la majorité des tâches et augmente significativement la précision (+1%). Une tuile complète est ainsi traitée en 4 minutes sur une NVIDIA Tesla K20c avec un pas de 32px et moins de 20 secondes avec un pas de 128px. Le temps d'inférence est doublé dans le cas des deux SegNet parallèles.

Transfert de connaissances. Comme détaillé dans le Tableau 1, le modèle réalise sa meilleure performance lorsque le taux d'apprentissage de l'encodeur est relativement faible. Ceci renforce l'idée que des filtres convolutifs génériques donnent les meilleurs résultats lorsqu'il est possible de laisser l'optimisation les spécialiser sur une tâche particulière. Cependant, il est important de souligner qu'une variabilité trop grande induit un risque de surapprentissage. Ainsi, il est possible d'utiliser le taux d'apprentissage des paramètres pré-entraînés comme régularisation lors de l'optimisation. Ces résultats sont similaires aux conclusions de [23] et aux observations générales de [34] concernant le transfert de connaissances.

Fusion de données par correction résiduelle. La fusion naïve des cartes de prédiction par simple moyenne n'améliore les résultats que par une faible marge (+0,3–0,4%). La fusion par apprentissage multiplie par deux ce gain. Comme illustré par la Figure 5, celle-ci combine les deux cartes de prédiction de façon plus pertinente, en exploitant leur complémentarité. Ainsi, la correction résiduelle donne des résultats visuellement plus cohérents. En particulier, l'image IRRV est plus efficace pour la prédiction des véhicules que notre image composite. En revanche, l'image composite a l'avantage du NDSM et du NDVI pour distinguer la végétation et discriminer entre arbre et herbe. Ainsi, le réseau de fusion donne plus de poids à l'image optique pour les voitures et plus de poids à l'image composite pour la végétation.
Cette approche de fusion améliore nos résultats sur le jeu de données ISPRS Vaihingen 2D Labeling Challenge jusqu'à 89,8% ("ONE_7", cf. Tableau 3). Les scores F1 sont significativement améliorés sur les bâtiments et la végétation grâce à l'apport du DSM et du NDVI. Cependant, bien

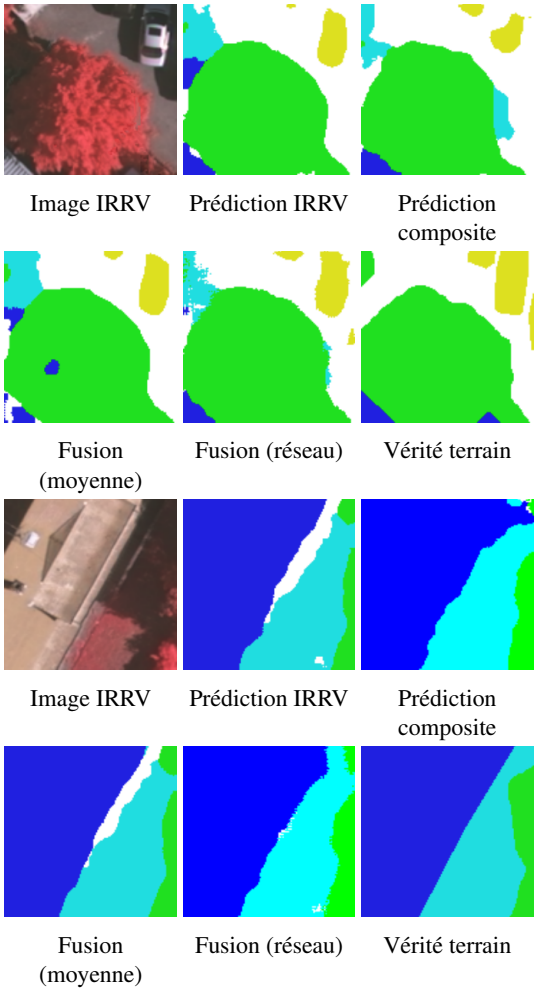


FIGURE 5 – Effets de la fusion de données sur différentes cartes de prédiction.

que compétitif, le score F1 des véhicules est nettement dégradé. Le décalage spatial des véhicules mobiles dans le (N)DSM et l'image optique en est la cause la plus probable.

5 Conclusion

Nous avons étudié l'utilisation des FCN pour la cartographie sémantique à partir d'images aériennes en zone urbaine. En particulier, nous avons montré qu'une architecture encodeur-décodeur comme SegNet s'adaptait aisément des images naturelles aux images de télédétection. Cela renforce l'idée que les banques de filtres convolutifs apprises sur une grande variété d'images peuvent servir dans de très nombreuses tâches visuelles. Nous avons également introduit un module de correction résiduelle pour la fusion de données tardive. Nous avons montré comment ce module neuronal peut estimer et corriger de légères erreurs dans des cartes de prédictions issues de sources hétérogènes. Enfin, nous avons démontré la pertinence de ces apports en les validant sur le jeu de données ISPRS Vaihingen, sur lequel nous avons amélioré l'état-de-l'art de 1% par rapport à des méthodes équivalentes sans données ad-

ditionnelles.

Remerciements. Les données aériennes sur Vaihingen ont été fournies par la Société Allemande de Photogrammétrie, Télédétection et Géoinformation (DGPF) [8] : `http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html`.
Les travaux de Nicolas Audebert sont financés par le projet de recherche NAOMI issu de la collaboration entre l'ONERA et Total. Les auteurs remercient l'Agence Nationale de la Recherche (ANR) pour leur soutien dans le cadre du projet ANR-13-JS02-0005-01 (Asterix).

Références

[1] Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip Torr. Higher Order Conditional Random Fields in Deep Neural Networks. *arXiv :1511.08119 [cs]*, November 2015.

[2] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-scale Deep Networks. In *Computer Vision – ACCV 2016*, pages 180–196. Springer, Cham, November 2016.

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet : A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *arXiv :1511.00561 [cs]*, November 2015.

[4] Alexandre Boulch. DAG of convolutional networks for semantic labeling. Technical report, Office national d'études et de recherches aérospatiales, 2015.

[5] M. Campos-Taberner, A. Romero-Soriano, C. Gatta, G. Camps-Valls, A. Lagrange, B. Le Saux, A. Beaupère, A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, M. Ferecatu, M. Shimoni, G. Moser, and D. Tuia. Processing of Extremely High-Resolution LiDAR and RGB Data : Outcome of the 2015 IEEE GRSS Data Fusion Contest Part A : 2-D Contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, PP(99) :1–13, 2016.

[6] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the Devil in the Details : Delving Deep into Convolutional Nets. In *Proceedings of the British Machine Vision Conference*, pages 6.1–6.12. British Machine Vision Association, 2014.

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan Yuille. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In *Proceedings of the International Conference on Learning Representations*, May 2015.

[8] M. Cramer. The DGPF test on digital aerial camera evaluation – overview and test design. *Photogrammetrie – Fernerkundung – Geoinformation*, 2 :73–82, 2010.

[9] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust RGB-D object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 681–687, September 2015.

[10] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes Challenge : A Retrospective. *International Journal of Computer Vision*, 111(1) :98–136, June 2014.

[11] Markus Gerke. Use of the Stair Vision Library within the ISPRS 2d Semantic Labeling Benchmark (Vaihingen). Technical report, International Institute for Geo-Information Science and Earth Observation, 2015.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers : Surpassing Human-Level Performance on ImageNet Classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[14] Sergey Ioffe and Christian Szegedy. Batch Normalization : Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 448–456, 2015.

[15] Adrien Lagrange, Bertrand Le Saux, Anne Beaupère, Alexandre Boulch, Adrien Chan-Hon-Tong, Stéphane Herbin, Hicham Randrianarivo, and Marin Ferecatu. Benchmarking classification of earth-observation data : From learning explicit features to convolutional networks. In *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 4173–4176, July 2015.

[16] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, April 2015.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO : Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, number 8693 in Lecture Notes in Computer Science, pages 740–755. Springer International Publishing, September 2014. DOI : 10.1007/978-3-319-10602-1_48.

[18] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, June 2015.

[19] Dimitrios Marmanis, Konrad Schindler, Jan Dirk Wegner, Silvano Galliani, Mihai Datcu, and Uwe Stilla. Classification With an Edge : Improving Semantic Image Segmentation with Boundary Detection. *arXiv :1612.01337 [cs]*, December 2016. arXiv : 1612.01337.

[20] Dimitrios Marmanis, Jan Dirk Wegner, Silvano Galliani, Konrad Schindler, Mihai Datcu, and Uwe Stilla. Semantic Segmentation of Aerial Images with an Ensemble of CNNs. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3 :473–480, 2016.

[21] Volodymyr Mnih and Geoffrey E. Hinton. Learning to Detect Roads in High-Resolution Aerial Images. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, number 6316 in Lecture Notes in Computer Science, pages 210–223. Springer Berlin Heidelberg, September 2010.

[22] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[23] Keiller Nogueira, Otávio A. B. Penatti Otávio A. B. Penatti, and Jefersson A. Dos Santos. Towards Better Exploiting Convolutional Neural Networks for Remote Sensing Scene Classification. *arXiv :1602.01517 [cs]*, February 2016.

[24] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning Deconvolution Network for Semantic Segmentation. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1520–1528, 2015.

[25] Sakrapee Paisitkriangkrai, Jamie Sherrah, Pranam Janney, and Anton Van Den Hengel. Effective semantic pixel labelling with convolutional networks and Conditional Random Fields. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 36–43, June 2015.

[26] Otávio A. B. Penatti Otávio A. B. Penatti, Keiller Nogueira, and Jefersson A. Santos. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains ? In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 44–51, June 2015.

[27] Nguyen Tien Quang, Nguyen Thi Thuy, Dinh Viet Sang, and Huynh Thi Thanh Binh. An Efficient Framework for Pixel-wise Building Segmentation from Aerial Images. In *Proceedings of the Sixth International Symposium on Information and Communication Technology*, page 43. ACM, 2015.

[28] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf. The ISPRS benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1 :3, 2012.

[29] Jamie Sherrah. Fully Convolutional Networks for Dense Semantic Labelling of High-Resolution Aerial Imagery. *arXiv :1606.02585 [cs]*, June 2016.

[30] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv :1409.1556 [cs]*, September 2014.

[31] M. Volpi and D. Tuia. Dense Semantic Labeling of Subdecimeter Resolution Images With Convolutional Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2) :881–893, 2017.

[32] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. High-performance Semantic Segmentation Using Very Deep Fully Convolutional Networks. *arXiv :1604.04339 [cs]*, April 2016.

[33] Zhicheng Yan, Hao Zhang, Yangqing Jia, Thomas Breuel, and Yizhou Yu. Combining the Best of Convolutional Layers and Recurrent Layers : A Hybrid Network for Semantic Segmentation. *arXiv :1603.04871 [cs]*, March 2016.

[34] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks ? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014.

[35] Fisher Yu and Vladlen Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. In *Proceedings of the International Conference on Learning Representations*, November 2015.

[36] Junbo Zhao, Michael Mathieu, Ross Goroshin, and Yann LeCun. Stacked What-Where Auto-encoders. In *Proceedings of the International Conference on Learning Representations*, June 2015.

[37] Wenzhi Zhao and Shihong Du. Learning multiscale and deep representations for classifying remotely sensed imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 113 :155–165, March 2016.

[38] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vineet Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. Conditional Random Fields as Recurrent Neural Networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1529–1537, 2015.